International Journal of Machine Consciousness
Vol. 2, No. 2 (2010) 343–361

DOI: 10.1142/S1793843010000539

World Scientific
www.worldscientific.com

EMOTIONS AND EMPATHY: A BRIDGE BETWEEN NATURE AND SOCIETY?*

RODRIGO VENTURA

Institute for Systems and Robotics,
Instituto Superior Técnico,
Av. Rovisco Pais, 1,
Lisbon, Portugal
rodrigo.ventura@isr.ist.utl.pt

For over a decade neuroscience has uncovered that appropriate decision-making in daily life decisions results from a strong interplay between cognition and covert biases produced by emotional processes. This interplay is particularly important in social contexts: lesions in the pathways supporting these processes provoke serious impairments on social behavior. One important mechanism in social contexts is empathy, fundamental for appropriate social behavior. This paper presents arguments supporting this connection between cognition and emotion, in individual as well as in social contexts. The central claim of this paper is that biologically inspired cognitive architectures ought to include these mechanisms. A taxonomy of computational models addressing emotions is presented, together with a brief survey of the research published in this area. The Prisoner Dilemma game is used as a case study exposing the trade-off between individual rationality and cooperative behavior. Experiments using a simple implementation of empathy and emotion expression, employing an Iterated Prisoner Dilemma setup, illustrate the emergence of a cooperative behavior mutually beneficial for both players.

Keywords: Emotions; empathy; cognitive architectures; Iterated Prisoner Dilemma.

1. Introduction

This paper addresses the role of emotions in cognition, in particular in what concerns its function in social behavior. According to the theory of evolution, all species are essentially survivors. Their actions are ultimately driven by individual survivability. However, individuals of some species interact socially in such a way the outcome is mutually beneficial [Axelrod and Hamilton, 1981; Trivers, 1971]. In many cases, this demands for letting go short-term individual benefits, in exchange of long-term mutual ones, even if the individual benefit supplants the latter. One example of this behavior is food sharing. In some cases, mutual benefits are a mere potential, or even nonexistent, as in the case of altruism. Such behavior contradicts, at least at a first glance, pure rationality. The term rationality is used along this paper in the

*This work was supported by the project FCT (ISR/IST plurianual funding) through the PIDDAC Program funds.

classical, utilitarian, decision theoretic sense of maximization of the expected utility [von Neumann and Morgenstern, 1944]. A broader view of rationality, encompassing emotions in particular, can be found, for instance, in [de Sousa, 1987]. In this paper, we limit our analysis to social contexts where individuals do have a choice of whether to cooperate, thus excluding hard-wired social behaviors such as the ones found in ant colonies.

The Prisoner Dilemma (PD) game is a paradigmatic case contrasting individual and cooperative behavior [Poundstone, 1993]. The classic description of this problem follows[a]:

> "*Two suspects are arrested by the police. The police have insufficient evidence for a conviction, and, having separated both prisoners, visit each of them to offer the same deal. If one testifies* (*defects from the other*) *for the prosecution against the other and the other remains silent* (*cooperates with the other*), *the betrayer goes free and the silent accomplice receives the full* 10*-year sentence. If both remain silent, both prisoners are sentenced to only six months in jail for a minor charge. If each betrays the other, each receives a five-year sentence. Each prisoner must choose to betray the other or to remain silent. Each one is assured that the other would not know about the betrayal before the end of the investigation. How should the prisoners act?*"

From an individual's standpoint, the rational choice would be to defect, regardless of the other's decision: if the other remains silent, defection is more beneficial (going free, against five years in prison), but if the other defects, defection is still more beneficial (five years against 10). Mutual defection is in fact the single Nash equilibrium point of this game. But if both defect, each one receives less than in the case when both cooperate. Although from an individual's point of view the best choice would be to defect, the cooperative option of both being silent is more (mutually) beneficial.

Neuroscientific evidence has uncovered that cooperative behavior depends critically on emotions [Adolphs, 2003; Damásio, 2003]. Evidence from humans playing the PD game have shown that the choice of cooperating rather than defecting is motivated by feelings of empathy with one another [Rilling *et al.*, 2002]. These feelings contribute to bias decision away from a pure rational choice, towards alternative, reciprocally altruistic options. It is now commonly accepted that emotions are a fundamental aspect of intelligent behavior, and in fact intelligence cannot be understood separately from emotions [Pessoa, 2008].

Emotional phenomena is in fact very broad in terms of their manifestations. Hudlicka [Hudlicka, 2009] distinguishes four different modalities: (1) behavioral/expressive, which concern expression and are visible by other persons (e.g., facial expressions), (2) somatic/neurophysiological, involving changes in the body state

[a] http://en.wikipedia.org/wiki/Prisoner_dilemma, retrieved 30-Mar-2010.

(e.g., heart rate), (3) cognitive/interpretative, concerning their implications in the cognitive processes in the brain, and (4) experiential/subjective, which relates to the first-person subjective experience of emotions. This paper will be primarily concerned with the cognitive/interpretative modality particularly in what concerns decision-making. However, it should be made clear that these four modalities do strongly interact.

This paper is organized as follows: Sec. 2 provides a brief historical perspective of intelligence understood as pure rationality, Sec. 3 reviews evidence on the role of emotions in decision-making processes of an individual, followed by Sec. 4 discussing its role in social contexts. Section 5 presents a survey of computational models of (and/or inspired by) emotional processes, followed by an illustrative example in Sec. 6 of a simple computational model for empathy in the iterated version of the Prisoner Dilemma game. Section 7 wraps up the paper with some conclusions.

2. On Rationality and Intelligence

People usually say "don't get emotional over this matter" in a manner of warning that emotions threaten to get in the way of the sound analysis of a situation. In fact, Western culture has been dominated by a Cartesian view of intelligence as dispassionate reasoning, happening in the realm of a disembodied mind. And on the contrary, emotions are viewed as something pertaining to the body, hence outside of the realm of reason. Intelligence and emotions are thus two things living in different, contradictory worlds.

The common sense idea of rationality opposing emotions can be traced back to the Greeks. For instance, Plato sustained an everlasting struggle between reason and emotion in our minds, with each one reaching for dominance over the other [Lyons, 1999]. This dualistic view lies behind the assumption that, if human level intelligence is sought, one should focus exclusively on rationality, factoring out the emotional. Intelligence has been understood as a synonym of reason.

During the first decades after the emergence of Artificial Intelligence (AI) as a field, this rational view of intelligence has been largely dominant. Despite many successes accomplished by the field, general intelligence constitutes, still, largely an open issue [Nilsson, 2005]. We argue on the assumption that one important missing link in the understanding of general intelligence is the role of emotions. The factorization of mental activity into reason and emotions, which has been silently assumed in many approaches to AI, has been questioned by neuroscientific evidence [Damásio, 1994; Pessoa, 2008]. The next section discusses some of these findings.

3. On Emotions and Decision-making

Damásio, among other researchers, have performed extensive studies on the role of emotion in decision-making, focusing on their neural correlates [Damásio *et al.*, 1991; Bechara *et al.*, 1997]. Although the modulatory effects that emotional phenomena

induces on mental activity have been thoroughly studied (e.g., attention focus, memory retrieval, etc.), he sustained that emotions are an integral part of decision-making processes. Moreover, he stresses that these mechanisms are founded on the body, and thus mind and body make an indivisible whole. This contrasts with the dualistic mind-body view of Descartes, thus motivating the name of his book "Descartes' Error" [Damásio, 1994]. This view is founded on his Somatic Marker Hypothesis (SMH), according to which mental imagery is associated with internal representations of body states [Damásio *et al.*, 1991; Damásio, 1994]. In certain situations (e.g., stressful), the brain associates mental imagery related with a situation with the alterations of the body state representations, induced by the emotional state. The associations thus formed can be re-enacted later, when the subject is experiencing a similar situation, or even when considering that situation as a possible consequence of a course of action. This re-enactment occurs using the same brain mechanisms as the ones prompting the body-state alterations following the emotion. This brain zone is the amygdala, serving as a central hub involving practically all emotion processes in the brain.

The implications of these processes in decision-making comes from a set of projections from the amygdala to the prefrontal cortex, where most high-level, cognitive processes are believed to occur (reasoning, planning, working memory, and so on). The function of these projections from the amygdala to the prefrontal cortex was studied by Damásio. It is from studies of patients with lesions in these projections that most evidence supporting the SMH comes from. Patients with these lesions behaved otherwise normally, except when facing certain decisions. The cognitive capabilities, as evaluated by I.Q. tests, turned out to be within normal. Damásio describes a particular case of such a patient that, when faced with the need to schedule his next meeting with him, the patient was unable to do so in useful time. He pondered endlessly the pros and cons of each possible option. Other reported consequences are the inability to make reasonable financial investment decisions, and difficulties in initiating a loving relationship. These patients usually lose their jobs, and marriages often dissolve. This suggests that the most practical, daily forms of decision-making depend critically on emotional mechanisms in the brain [Damásio, 1994].

These findings support the hypothesis that it is in situations where future outcomes are more uncertain, in the sense of being hard to predict in detail, that emotion processes provide a crucial contribution for appropriate decision-making.

Aaron Sloman has contested the claim that emotions are a prerequisite for intelligent behavior, on the following grounds [Sloman 1998; 1999]: the lesions studied by Damásio impair both (1) emotions and (2) decision-making, from which one cannot infer that the effect (1) implies effect (2). This does not contradict, however, the hypothesis that there are processes in the brain, which support emotions, that are essential for appropriate decision-making. Moreover, the effects of these processes in decision-making are known to take place even if no emotion is felt by the subject.

4. On Emotions and Social Cognition

One prime example of dynamically changing environments are social contexts. Here, the pairing between actions and reward expectancy may change dramatically with time. The Iterated Prisoner Dilemma (IPD) game is a paradigmatic example of such a situation. The IPD is an iterated version of the Prisoner Dilemma game (described in Sec. 1), where each player has access to the previous turns, and the payoffs are monetary, specified in Table 1.

From an individual point of view, each player maximizes his payoff by defecting. For example (Table 1(b)), if they both defect, the payoff is \$1 for each, while mutual cooperation yields \$2 to both of them. However, when one cooperates and the other defects, the former gets \$0, while the defector gains \$3. The dynamics of this game comes from the fact that, if one of the agents switches strategy (e.g., from cooperation to defection), the opponent has to rapidly adjust his own strategy.

With the use of brain imaging techniques, subjects have been scanned while playing IPD games. It was found that brain regions implicated in the SMH mechanism bias decision-making towards cooperative behavior in this game [Rilling *et al.*, 2002; Adolphs, 2003]. In particular, the mechanism of re-enactment of the body state representation seems to be crucial for subjects to cooperate. Failure to do so, as can be observed in patients with specific brain lesions, lead subjects to defect, preferring immediate rewards, in exchange for the long-term benefits of mutual cooperation. Interestingly, after the experiments, control subjects reported that mutual cooperation was the most personally satisfying outcome, while defection provoked feelings of guilt.

One of the serious consequences of the lesions studied by Damásio concerned social behavior. From his studies of patients with lesions affecting emotional mechanisms, he reported that they lose the ability to make appropriate decisions under uncertainty. For instance, they showed severe impairment in empathy, as well as maintaining personal trust, adequate social behavior, maintaining marriage and a healthy relationship with the offspring. But strikingly, intellectual capabilities remained intact, as they were (verbally) aware of the social rules they themselves break [Damásio, 2003].

Table 1. Payoffs for the Iterated Prisoner Dilemma game: (a) canonical payoffs table, where $T > R > P > S$ and $2R > T + S$ [Nowak and Sigmund, 1990]; (b) example payoffs. Actions: C = cooperate, D = defect. Payoffs in the form P_1/P_2 where P_1 and P_2 are the payoffs for players 1 and 2.

(a) Canonical payoffs

player 1	player 2	
	C	D
C	R/R	S/T
D	T/S	P/P

(b) Example payoffs

player 1	player 2	
	C	D
C	\$2/\$2	\$0/\$3
D	\$3/\$0	\$1/\$1

Empathy is an important mechanism for social interaction. Brain imaging studies have revealed that empathy is based on changing one's internal body representation, by the replication of the feelings of others. One study has shown that this representation is more intense when imitating a facial expression than when observing one [Carr *et al.*, 2003]. Another study provided evidence that we understand the pain of others by instantiating it internally in our brain [Jackson *et al.*, 2005]. The importance of this internal re-enactment of feelings of others is corroborated by evidence revealing that lesions in the amygdala compromise more the perception of social emotions from faces, than simple emotions [Baron-Cohen and Tranel, 2002]. This re-enactment of internal body states after the observation of the same states in others resonates nicely with the discovery of the mirror neurons [Carr *et al.*, 2003]: these neurons are both active while performing a goal-directed action, and while watching someone else performing the same action [Gallese *et al.*, 1996].

In summary, neuroscientific evidence has supported the hypothesis that appropriate social behavior depends critically on emotional mechanisms in the brain [Rilling *et al.*, 2008]. One reason for this may be the uncertain nature of the decisions involved in social contexts. When uncertainty (or unpredictability) in future outcomes is present, it is hard to draw conclusions of the benefits of these outcomes, based on rational principles alone. Thus, emotions emerge as an alternative mechanism for making decisions. Thus, it is under uncertainty that emotional mechanisms are more relevant for decision-making.

5. On Computational Models of Emotions

The main goal of this section is to provide a taxonomy of computational models addressing emotions, together with a brief survey of recent research. A broad variety of research is covered, ranging from interactive systems to emotion-based cognitive architectures.

5.1. *Early approaches*

The idea of using emotion mechanisms in artificial systems was first proposed by Herbert Simon in 1967 [Simon, 1967]. In his paper, Simon considers systems with multiple goals. When faced with real-time systems, where the survival of the system depends on its response time in certain situations, an *interrupt system* capable of interrupting current processing in order to attend to a real-time solicitation is considered by Simon as an emotional behavior mechanism. A few decades afterwards, Aaron Sloman and Monica Croucher sketched a complex architecture of the mind, in which emotions play an important role. In a similar fashion as Simon, emotions are taken to play the role of interrupting current processing in order to cope with the vicissitudes of a changing and partly unpredictable environment. Sloman and Croucher conjecture in their paper from 1981 that "interruptions, disturbances and departures from rationality which characterize emotions are a natural consequence of

the sorts of mechanisms required by the constraints on the design of intelligent systems" [Sloman and Croucher, 1981].

Both Simon and Sloman approached emotions as mechanisms of a larger system. They claimed that a cognitive system complex enough to exhibit intelligent behavior at a level similar to the human one ought to incorporate emotion-like mechanisms. Note that the focus is not on specific human emotions, but rather on the mechanisms involved in emotion processing.

An alternative approach is possible, though: the design of a system endowed with representations and mechanisms closely based on human emotions. Early works on this line of research include the ones of Jaap Swagerman, based on the emotion theories of Nico Frijda [1986]. In [Dyer, 1987] Michael Dyer reviews previous computer models that exhibit comprehension and/or generation of emotional behavior.

5.2. *Taxonomy*

In the taxonomy proposed here, computational approaches to emotions are first divided in two main areas, designated here as focused on *internal* and *external* manifestations of emotions. This division concerns whether the design of the artifact is more focused on the internal implications of emotions on cognitive processes, or on the interactive/communicative aspects of emotions.

Alternative taxonomies of computational models can be found in the literature. For instance, Hudlicka has proposed a categorization of the approaches among emotion *generation*, addressing how emotions arise in a given situation, and emotion *effects*. In this latter category, Hudlicka distinguishes between visible, behavioral effects, and the less visible influence on attention, perception, and cognition [Hudlicka, 2008]. These two subcategories can be mapped onto the division of the approaches among internal and external manifestations proposed in this paper.

The division proposed here, among internal and external manifestations of emotions, corresponds roughly to the modalities (1) behavioral/expressive and (2) cognitive/interpretative referred in Sec. 1. These two approaches are not hermetic, as they are both inspired in the same integrated phenomena of biological emotions.

Within each of these two areas, the proposed taxonomy further divides research in a set of subareas. The criterion guiding the choice of subareas is based on the research goal stated by the authors. For instance, the architectures subarea concerns the construction of architectures, while robotics includes the implementation in a real (or simulated) robot. The adopted taxonomy, organized in two levels of details, follows. A more in-depth review of the cited literature can be found in the thesis [Ventura, 2008]. Here we will limit ourselves to a survey of the prominent approaches.

5.2.1. *Internal manifestations of emotions*

Architectures. Research in this area aims at a generic agent architecture where the internal mechanisms of emotions play a prominent role. Examples of this area include: Sloman's CogAffect architecture [Sloman, 1998], based on viewing emotions

as an alarm system; Toda's Emotional Fungus-eater [Toda, 1982], where emotions are viewed as *urges* (motivational subroutines), which were further developed by Aubé [Aubé, 1998], adding the concept of emotions as commitment operators involving two or more agents; Pfeifer's FEELER model [Pfeifer, 1994] driven by a set of production rules, based on an emotions taxonomy proposed by the psychologist Bernard Weiner in 1982; Botelho's Salt & Pepper architecture [Botelho *et al.*, 2004] where emotion signals and responses are generated by an affective appraisal process based on a production system; Burt's agent architecture, where emotions function as a scheme for managing resources in a three-layered architecture; Velasquez's Cathexis architecture [Velásquez, 1998], founded on Minsky's Society of Mind [Minsky, 1988]; and Staller's TABASCO architecture [Staller and Petta, 1998], a three-layered architecture (conceptual, schematic, and sensory) employing an appraisal approach.

Robotics. The goal in this area is the construction of mobile robots whose behavior is determined by emotional components in the architectures. Although the robotics area can be seen as an application of research on the architectures above, we included here the research work in which both one or more robots are involved (either physical or in simulation), together with its kinematic constraints, and when the proposed approach is somehow dedicated to robots. Examples include: Beaudoin's NML1 [Beaudoin, 1994] and Wright's MINDER1 [Wright, 1997] agent implementations, based on Sloman's architecture; Cañamero's agents [Cañamero, 1997] living in a grid-world, based on Minsky's Society of Mind paradigm [Minsky, 1988]; Scheutz's architecture [Scheutz, 2001] where several agents evolve in a 2D environment, driven by a schema-based controller [Arkin, 1989]; Gadanho's architecture [Gadanho and Hallam, 2002] targeting Khepera robots, combining reinforcement learning with an emotional system; Gmytrasiewicz's theoretic approach to emotions [Doshi and Gmytrasiewicz, 2004]; and Morgado's signal processing approach [Morgado and Gaspar, 2005], modeling the dynamics of variables, such as achievement potential, and achievement conductance.

Emotions modeling. This area aims at the creation of models of mechanisms of emotions, not necessarily biologically-inspired. Examples include: Arzi-Gonczarowski's formal modeling [Arzi-Gonczarowski, 2000] based on mathematical category theory; Gratch's approach [Gratch, 2000] based on cognitive appraisal theories, modeling the influence of emotions on the planning process of an autonomous agent; and Wilson's Artificial Emotion Engine model [Wilson, 2000] employing Eysenck's model of personality traits.

Cognitive modeling. The research surveyed here addresses computational models of emotional mechanisms of the brain. The Emotion Modeling subarea is here distinguished from Cognitive Modeling one by the object being modeled: the latter aims at modeling cognitive mechanisms in humans, by the means of computational models, while the former is here understood in the context of (abstract) artificial models of emotions. Examples include: Balkenius's computational model of emotional learning and processing [Balkenius and Morén, 2001], where several brain

areas are explicitly modeled at a functional level, rather than at a neural level; Dörner model of human action regulation (Psi-model) [Dörner and Starker, 2004], integrating cognition, motivation, and emotion; Fellous' model [Fellous, 2004] viewing emotions as dynamic patterns of neuromodulation, rather than patterns of neural activity as it is traditionally done; Hudlicka's computational cognitive-affective architecture (MAMID) [Hudlicka, 2004] where the underlying idea is that affective states, together with personality traits (individual differences), manipulate a series of architectural parameters, such as the processing speed and capacity of a set of cognitive modules; Lisetti's neural network, modeling the human Autonomous Nervous System (ANS), capturing emotion processing at the physiological level [Lisetti, 1998]; Almeida's physiological model of the body [de Almeida *et al.*, 2004], at the organ level.

5.2.2. *External manifestations of emotions*

Believable agents. The goal in this area is to build interactive agents seeking suspension of disbelief with the user. Examples include: the Oz project at CMU addressing the construction of several interactive believable agents [Bates *et al.*, 1994; Reilly, 1996]; Blumberg's AlphaWolves project, consisting of a social environment of a pack of virtual gray wolves [Tomlinson *et al.*, 2002]; Elliott's Affective Reasoner (AR) based on OCC theory, with the goal of simulating several aspects of emotion processing in a multi-agent setup [Elliott, 1992]; Numaoka's system targeted for the design of a personal assistant in a virtual reality setup [Numaoka, 1998]; Martinho's virtual reality installation for the Expo'98 World Fair, consisting of a pair of virtual dolphins interacting with the audience [Martinho *et al.*, 2000]; and Aylett's virtual Teletubbies (based on the well-known homonymous TV series for children), targeting collective behaviors of virtual sheep [Aylett, 1999].

Affective Human-Computer Interfaces (HCI). Traditional HCI is based on interaction with the user on a rational basis, while affective HCI focuses on affective interaction among users and computers. This includes two aspects: computers recognizing affective states of users, and computers expressing emotional states in a believable way. Affective Computing, a term coined by Picard in 1995 [Picard, 1995], proposes a shift on the way humans interact with machines, from a traditional, rational and deterministic basis, towards an interaction conveying affective content. The expression *affective computing* has since then gained a broader application, being often used to denote any computational model of emotions. Thus, for the sake of clarity, we have decided to categorize the following approaches as *Affective HCI.* To attain believability of the expressed emotions, some form of emotion modeling is required, using for instance one of the approaches referred in previous paragraphs. Examples of affective HCI include Picard's research on techniques for measuring emotions [Vyzas and Picard 1998; Picard *et al.*, 2001], together with innovative applications [Healey *et al.*, 1998; Picard and Scheirer, 2001]; Cynthia's Kismet [Breazeal, 2002] robot head, capable of being sensitive and expressing a broad range of

emotions; Moshkina's AuRA robot architecture [Arkin and Balch, 1997], modeling personality traits, attitudes, moods, and emotions; and Conati's probabilistic model of a user while interacting with educational games [Conati and Zhou, 2004].

5.3. *Emotion-based agents*

Ventura *et al.* proposed in 1998 an emotion-based agent architecture [Ventura and Pinto-Ferreira, 1998; Ventura *et al.*, 1998] inspired in Damásio's Somatic Marker Hypothesis (SMH) [Damásio *et al.*, 1991; Damásio, 1994]. This architecture is founded on the principle that stimuli is processed internally by two layers with different degrees of complexity and accuracy. These layers correspond to a (1) *perceptual layer*, providing a reactive, quick response to stimuli, and thus giving a primordial meaning to stimuli eliciting a response, and a (2) *cognitive layer*, representing stimuli with complex, high-dimensionality representations. This architecture was further developed and formalized in [Ventura and Pinto-Ferreira, 2007; 2008].

Of particular relevance for social contexts, is the work of Maçãs *et al.*, augmenting the model with an extra layer: a symbolic layer [Maçãs and Custódio, 2003]. This extension was implemented in a market environment, where products are exchanged for money among agents. The agents seek survival, as well as the maximization of profit from selling goods. There is explicit communication among agents, in which the symbolic layer plays a central role. In this framework, the cognitive and the symbolic layers distinguish themselves in the fact that, while the former is focused on individual behavior, the latter accounts for social issues. Social interaction enables an agent not only to take into account its own experience, but also the experience of others. This is done in a similar fashion than empathy: "When a sequence ends because of another agent change of expression, it is evaluated as if it was the own agent. This is the process of gathering and storing others' experiences." [Maçãs and Custódio, 2003].

6. Illustrative Example

As an illustration that reciprocally altruistic behavior can result from empathy, a simple example was devised using the IPD domain. The motivation for using the IPD game was that it presents a trade-off between individualistic and cooperative behavior, in such a way that cooperative behavior is more mutually beneficial, than the individually rational choice (Sec. 1).

Consider two agents playing the Iterated Prisoner Dilemma game, each one with its own strategy. In each turn, both agents are asked to perform one of two possible actions: to cooperate (C), or to defect (D). The payoffs each agent receives at each turn are specified in Table 1(b). The turns are iterated a specified amount of times, and the performance of each player is assessed by the sum of the obtained payoffs.

6.1. *Rational agent*

The strategy of the rational agent is based on the maximization of the expected utility principle. The agent estimates the expected utility of each option (C or D), and chooses the one maximizing the expected utility of the outcome. The expected utility is computed with a moving average of the past rewards with a fixed window. This is performed independently for each one of the actions.

Formally, we denote the history of the actions up to (discrete) time t by the vector A_t, defined by

$$A_t = [a(1) \cdots a(t)] \tag{1}$$

and the corresponding rewards[b] by the vector R_t, defined by

$$R_t = [r(1) \cdots r(t)], \tag{2}$$

where $a(t) \in \{D, C\}$ and $r(t) \in \mathbb{R}$ are the agent action and the resulting reward after turn t. Splitting these sequences with respect to the action performed by the agent we can write the ordered sequence of indices for which each action $a \in \{D, C\}$ as

$$I_t^a = [t_1^a \cdots t_{n_a}^a] \quad \text{such that } a(t_i^a) = a, \quad i = 1, \ldots, n_a, \tag{3}$$

where $t_i^a < t_{i+1}^a$ for all i. The moving average to estimate the expected utility of action a can then be expressed by

$$EU_a(t) = \frac{1}{L} \sum_{k=0}^{L-1} r(t_{n_a-k}^a), \tag{4}$$

where L is the window size. The rational strategy boils down to the maximization of this expected utility over the possible actions

$$a_{t+1}^{\text{rat}} = \underset{a \in \{D,C\}}{\arg\max}\ EU_a(t). \tag{5}$$

6.2. *Emotion-empathic agent*

The design of the emotion-empathic agent follows two simple principles, inspired by the findings reviewed in Sec. 4: (i) each agent (faithfully) expresses an emotional response corresponding to the difference between the received reward and the expected utility for the performed action, and (ii) the agent decision takes into account not only the expected utility, but also the expected emotional response of the opponent (empathy). The first principle follows from the evidence that emotion expression is fundamentally innate and faithful.[c] Empathy, in the sense of feeling what the other is feeling, is realized by the second principle: an agent's actions are not only determined by its individual payoff, but also by the expected empathic feeling.

[b] The terms payoff and reward will be used interchangeably throughout the rest of the paper.

[c] Exceptions exist, such as in expression containment and in dramatic play, but demand effort (and training) of some sort by the subject.

This is accomplished by considering a weighed sum of the expected reward with the expected emotion expressed by the opponent.

For the sake of simplicity, emotion is here modeled as a scalar value of valence (positive is good, negative is bad, zero is neutral). Under the assumption that agents have a preference for positive states, we can map this valence to a utility value. This is assumedly an over-simplification of affective phenomena, but for the purposes of this illustration it is sufficient.

This agent is implemented as an extension to the rational agent above, with the following modifications. First, the capability of expressing the emotional response. Let us denote the emotion expressed by one agent after performing action a in turn t by

$$em(t) = r(t) - EU_a(t-1). \tag{6}$$

Following principle (i), the expressed emotion is the difference between the received payoff and the expected utility of the performed action a. All of the formalism introduced so far concerns one of the agents involved in the game. We will denote the variables for the other agent with a line over the variable, e.g., $\overline{em}(t)$ is the emotion expressed by the other agent at t. An emotion-empathic agent computes the expected emotional response of the other agent by performing a moving average over its emotional responses

$$\overline{EEM}_a(t) = \frac{1}{M}\sum_{k=0}^{M-1} \overline{em}(\overline{t^a}_{n_a-k}), \tag{7}$$

where M is the window size. The emotion-empathic agent decision can then be formalized as follows, implementing the principle (ii) above:

$$a_{t+1}^{\text{emp}} = \arg\max_{a\in\{D,C\}} \left[EU_a(t) + \lambda \overline{EEM}_a(t)\right]. \tag{8}$$

The agent decision aims at the maximization of the expected utility, biased by the expected emotional response of the opponent. In other words, the empathy with the opponent's emotional expression (in expectation) exerts a bias on the rational decision.

From expression (8) one realizes that there is a trade-off between the complete selfishness of a rational agent (with $\lambda = 0$), where (8) degenerates into (5), and acting altruistically (for $\lambda \gg 0$).

Alternative approaches to IPD playing employing affective models can be found in the literature. In [Kim and Taber, 2004], for instance, a model using the ACT-R cognitive architecture [Anderson *et al.*, 2004] is used to experiment with the IPD game, modeling both affective and cognitive mechanisms.

6.3. *Experimental results and discussion*

The experimental setup we used to evaluate the performance of the proposed emotion-empathic agent consisted in running several IPD games and collecting

statistics of the payoffs received by each agent. The agent strategies considered were the following:

emp — the emotion-empathic agent, which expresses an emotional state defined in Eq. (6) and decides its action according to Eq. (8);
rat — the rational agent, employing strategy Eq. (5), with zero emotional expression;
rat+ — like **rat** but expressing an emotion response according to Eq. (6);
t4t — the classic tit-for-tat strategy: action is the opponent's last move, while cooperating in the first turn;
t42t — the tit-for-two-tats strategy: action is *cooperate* unless the opponent's two last actions were to *defect*;
t4t+ and t42t — like **t4t** and **t42t** but expressing emotional responses according to (6);
rnd — a random agent, which actions are randomly drawn with equal probability.

For the **emp**, **rat** and **rat+** agents, a simple exploration strategy was employed to allow for a better estimation of the expected utility and emotion expression. This exploration strategy consists on performing an equally probable random action with (decaying) probability $p(t) = \gamma^{t-1}$ (for $t = 1, 2, \ldots$) with $0 < \gamma < 1$.

The results were collected after 100 games of 1000 steps each. The parameters used were $\lambda = 0.5, M = L = 10$, and $\gamma = 0.9$. For each game, the average payoff (accumulated payoff divided by the game length) was recorded. The results are presented as the average payoff over all games, for each combination of agents (Table 2).

From these results we highlight the following observations, all statistically significant with $p < 0.01$:

(1) Two emotion-empathic agents perform better than two rational agents; the average payoff is however lower than full cooperation (average payoff of 2), and in fact, each agent only converges to a cooperative strategy in about 75% of the games. This can be explained by the initial exploratory phase, since during this

Table 2. Average values for 100 games, of the average payoff for 1000 steps games. See text for the description of each agent strategy. For each agent combination, the upper number corresponds to the value for the row agent, while the lower one to the column agent.

emp	1.75	1.03	1.02	1.87	1.08	1.96	1.05	1.99
	1.75	1.08	1.02	1.87	1.07	1.94	1.03	0.52
rat		1.05	1.03	1.05	1.08	1.06	1.08	1.99
		1.05	1.03	1.04	1.07	1.04	1.06	0.51
t4t			1.07	2.00	2.00	2.00	2.00	1.50
			1.08	2.00	2.00	2.00	2.00	1.50
t42t				2.00	2.00	2.00	2.00	1.25
				2.00	2.00	2.00	2.00	2.00
	emp	rat+	rat	t4t+	t4t	t42t+	t42t	rnd

phase the *cooperate* actions is very likely to be paired with an opponent's *defect*, thus biasing the agent towards defection.

(2) The emotion expression in the **t4t+** and **t42t+** variants, when playing with an emotion-empathic agent, result extremely beneficial for both players (compare 1.87/1.87 of **emp/t4t+** with 1.08/1.07 of **emp/t4t**, and 1.96/1.94 of **emp/t42t +** with 1.05/1.03 of **emp/t42t**). This follows directly from the bias of the emotion-empathic agent towards cooperation, since defection of one agent causes lower values of emotion response by the opponent.

(3) Both the emotion-empathic and the rational agents outperform both the tit-for-tat and tit-for-two-tats when playing against the random agent. This is caused by the maximization of the expected utility principle, which leads these agents to converge to a defective strategy against an agent that randomly cooperates/defects (i.e., there is no practical advantage in cooperating).

(4) Any combination of the tit-for-tat and of the tit-for-two-tats agents, including emotion expressive variants, yields full cooperation all the time (payoff of 2), simply because in this strategies no agent has ever the initiative to defect.

The parameter λ is responsible for the trade-off between selfishness and altruism. What is the effect of this parameter in the results? Figure 1 plots the average payoff, in the same experimental conditions as before, in function of λ. This plot compares the average payoff (as defined above) for two configurations — two emotion-empathic agents (emp/emp), and one emotion-empathic agent against a rational one with emotion expression (emp/rat+) — with λ varying from 0 to 1.4. As λ increases, the average payoff of the two emotion-empathic agents also increase, converging to cooperative strategies for $\lambda > 0.5$. We can then conclude that in this context, altruism promotes cooperation. There is, however, a price to pay: when playing against a rational agent, it takes advantage of the opponent's altruism and converges to defection. The emotion-empathic agent does not bother cooperating, since the positive emotion expressed by the rational agent overrules its own disadvantageous

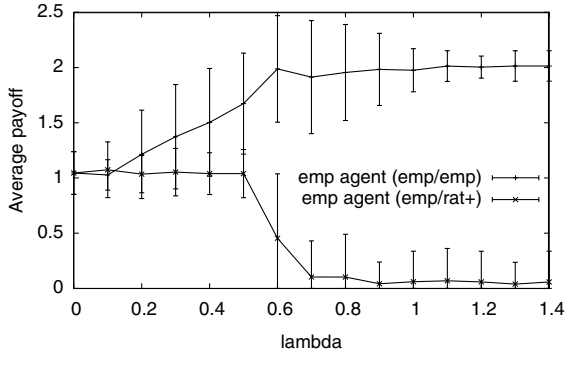

Fig. 1. Average payoff of an **emp** agent against another **emp** and against a **rat+** one, for a range of λ values. Error bars denote standard deviation.

rewards. There is however a range of values of λ, between about 0.2 and 0.5, for which the emotion-empathic strategy converges to a defection strategy when playing against a rational agent, and to a (frequent) cooperative strategy when playing against an emotion-empathic agent.

It is important to note that strategies based on the tit-for-tat and on the tit-for-two-tats are specifically designed for the IPD domain, in the sense that they rely on the structure of the payoff table (Fig. 1). Instead, the strategies **emp**, **rat** and **rat+** learn from scratch the relation between the payoffs and actions. It would be possible, in principle, to design payoff tables such that the latter strategies would perform equally well, while the former ones would fail.

7. Conclusions

The main purpose of this paper is to highlight the importance of emotion processes to intelligent behavior, from the design of artificial systems standpoint. There is empirical evidence that emotions are capable of biasing decisions away from the rational choice, but it turns out that this bias can be mutually beneficial from a social point of view (often in a different time scale).

Recent evidence sustaining the importance of emotional mechanisms in human decision-making was reviewed. The role of these mechanisms was discussed first at the individual level, and then in social contexts. In fact, emotions were found to be particularly important for appropriate social behavior.

A survey of computational models addressing emotions was performed, framed by a proposal taxonomy of the field. Both external and internal manifestations of emotions were covered, showing a wide range of research in these area for the past decades.

An illustrative implementation exploring the effects of emotional expression and empathy in the domain of the Iterated Prisoner Dilemma was presented, together with experimental results contrasting several strategies. These strategies included the rational choice, based on a plain maximization of expected utility, and the effect of biasing the rational choice taking into account the expected emotional response of the opponent. Results have shown that this bias contributes for a course of action more advantageous to both, than taking the rational choice.

From the perspective of the design of intelligent machines, we claim that inspiration from biology is a rich paradigm for advancing the field. In this line, research on biologically-inspired cognitive architectures should take into account that emotions are integrated in the very process of human intelligence.

Acknowledgments

The author would like to thank Eva Hudlicka for the insightful comments to an early draft of this paper.

References

Adolphs, R. [2003] "Cognitive neuroscience of human social behaviour," *Nature Reviews Neuroscience* **4**, 165–178.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C. and Qin, Y. [2004] "An integrated theory of the mind," *Psychological Review* **111**(4), 1036–1060.

Arkin, R. C. and Balch, T. [1997] "AuRA: Principles and practice in review," *Journal of Experimental and Theoretical Artificial Intelligence* **9**(2–3), 175–189.

Arkin, R. C. [1989] *Behavior-Based Robotics* (MIT Press).

Arzi-Gonczarowski, Z. [2000] "A categorization of autonomous action tendencies: The mathematics of emotions," in *Cybernetics and Systems 2000*, R. Trappl (ed.), Vol. 2 (Wien), pp. 683–688.

Aubé, M. [1998] "A commitment theory of emotions," *Emotional and Intelligent: The Tangled Knot of Cognition, Papers from the AAAI Fall Symposium, Technical Report FS-98-03* (AAAI Press, Menlo Park, California), pp. 13–18.

Axelrod, R. and Hamilton, W. D. [1981] "The evolution of cooperation," *Science* **211**, 1390–1396.

Aylett, R. [1999] "Emotion in behavioural architectures," *Workshop on Emotion-Based Agent Architectures (EBAA'99)*, pp. 10–12.

Balkenius, C. and Morén, J. [2001] "Emotional learning: A computational model of the amygdala," *Cybernetics and Systems: An International Journal* **32**, 611–636.

Baron-Cohen, R. A. S. and Tranel, D. [2002] "Impaired recognition of social emotions following amygdala damage," *Journal of Cognitive Neuroscience* **14**(8), 1264–1274.

Bates, J., Loyall, A. B. and Reilly, W. S. [1994] "An architecture for action, emotion, and social behavior," in *Artificial Social Systems* LNCS 830 (Springer), pp. 55–68.

Beaudoin, L. [1994] *Goal Processing in Autonomous Agents*, Ph.D. thesis, School of Computer Science, University of Birmingham.

Bechara, A., Damásio, H., Tranel, D. and Damásio, A. R. [1997] "Deciding advantageously before knowing the advantageous strategy," *Science* **275**, 1293–1295.

Botelho, L., Ramos, P. and Figueiredo, P. [2004] "Emotion eliciting in Salt & Pepper," *Cybernetics and Systems 2004*, in R. Trappl (ed.), (Wien), pp. 657–662.

Breazeal, C. [2002] "Regulation and entrainment in human-robot interaction," *The International Journal of Robotics Research* **21**(10), 883–902.

Cañamero, D. [1997] "Modeling motivations and emotions as a basis for intelligent behavior," *Proceedings of the First International Conference on Autonomous Agents*, pp. 148–155.

Carr, L., Iacoboni, M., Dubeau, M.-C., Mazziotta, J. C. and Lenzi, G. L. [2003] "Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas," *Proceedings of the National Academy of Sciences of the United States of America* **100**(9), 5497–5502.

Conati, C. and Zhou, X. [2004] "A probabilistic framework for recognizing and affecting emotions," *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, Papers from 2004 AAAI Spring Symposium, Technical Report SS-04-02* (AAAI Press, Menlo Park, California), pp. 13–16.

Damásio, A. R. [1994] *Descartes' Error: Emotion, Reason and the Human Brain* (Picador).

Damásio, A. R. [2003] *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain* (Harvest Books).

Damásio, A. R., Tranel, D. and Damásio, H. C. [1991] *Frontal Lobe Function and Dysfunction*, chap. Somatic Markers and the Guidance of Behavior: Theory and Preliminary Testing (Oxford University Press, NY), pp. 217–229.

de Almeida, L. B., da Silva, B. C. and Bazzan, A. L. C. [2004] "Towards a physiological model of emotions: First steps," *Architectures for Modeling Emotion: Cross-Disciplinary*

Foundations, Papers from 2004 AAAI Spring Symposium, Technical Report SS-04-02 (AAAI Press, Menlo Park, California), pp. 1–4.

de Sousa, R. [1987] *The Rationality of Emotion* (MIT Press).

Dörner, D. and Starker, U. [2004] "Should successful agents have emotions? The role of emotions in problem solving," *Proceedings of the Sixth International Conference on Cognitive Modeling* (Lawrence Earlbaum, Mahwah, NJ), pp. 344–345.

Doshi, P. and Gmytrasiewicz, P. [2004] "Towards affect-based approximations to rational planning: A decision-theoretic perspective to emotions," *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, Papers from 2004 AAAI Spring Symposium, Technical Report SS-04-02* (AAAI Press, Menlo Park, California), pp. 33–36.

Dyer, M. G. [1987] "Emotions and their computations: Three computer models," *Cognition and Emotion* **1**(3), 323–347.

Elliott, C. D. [1992] *The Affective Reasoner: A process model of emotions in a multi-agent system*, Ph.D. thesis, Northwestern University, Evanston, Illinois.

Fellous, J.-M. [2004] "From human emotions to robot emotions," *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, Papers from 2004 AAAI Spring Symposium, Technical Report SS-04-02* (AAAI Press, Menlo Park, California), pp. 37–47.

Gadanho, S. C. and Hallam, J. [2002] "Robot learning driven by emotions," *Adaptive Behavior* **9**(1), 42–64.

Gallese, V., Fadiga, L., Fogassi, L. and Rizzolatti, G. [1996] "Action recognition in the premotor cortex," *Brain* **119**, 593–609.

Gratch, J. [2000] "Émile: Marshalling passions in training and education," *Proceedings of the 4th International Conference on Autonomous Agents*, Barcelona, Spain, pp. 325–332.

Healey, J., Picard, R. and Dabek, F. [1998] "A new affect-perceiving interface and its application to personalized music selection," *Proceedings of the 1998 Workshop on Perceptual User Interfaces*, San Francisco, CA, pp. 4–6.

Hudlicka, E. [2004] "Two sides of appraisal: Implementing appraisal and its consequences within a cognitive architecture," in *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, Papers from 2004 AAAI Spring Symposium, Technical Report SS-04-02* (AAAI Press, Menlo Park, California), pp. 70–76.

Hudlicka, E. [2008] "What are we modeling when we model emotion?" *Proceedings of the AAAI Spring Symposium on "Emotion, Personality and Social Behavior,"* pp. 52–59.

Hudlicka, E. [2009] "Challenges in developing computational models of emotion and consciousness," *International Journal of Machine Consciousness* **1**(1), 131–153.

Jackson, P. L., Meltzoff, A. N. and Decety, J. [2005] "How do we perceive the pain of others? A window into the neural processes involved in empathy," *NeuroImage* **24**, 771–779.

Kim, S. and Taber, C. [2004] "A cognitive/affective model of strategic behavior-2-person repeated prisoner's dilemma game," *Proceedings of the Sixth International Conference on Cognitive Modeling* (Pittsburgh, PA), pp. 360–361.

Lisetti, C. L. [1998] "Emotion synthesis: Some research directions," *Emotional and Intelligent: The Tangled Knot of Cognition, Papers from the AAAI Fall Symposium, Technical Report FS-98-03* (AAAI Press, Menlo Park, California), pp. 109–115.

Lyons, W. [1999] "The philosophy of cognition and emotion," *Handbook of Cognition and Emotion* (Wiley), pp. 21–44.

Maçãs, M. and Custódio, L. [2003] "Multiple emotion-based agents using an extension of DARE architecture," *Informatica* **27**, 185–195.

Martinho, C., Paiva, A. and Gomes, M. [2000] "Emotions for a motion: Rapid development of believable pathematic agents in intelligent virtual environments," *Applied Artificial Intelligence* **14**(1), 33–68.

Minsky, M. [1988] *The Society of Mind* (Touchstone).

Morgado, L. and Gaspar, G. [2005] "Emotion-based adaptive reasoning for resource bounded agents," *Proceedings of AAMAS'05*, Utrecht, Netherlands, pp. 921−928.

Nilsson, N. J. [2005] "Human-level artificial intelligence? Be serious!" *AI Magazine* **26**(4), 68−75.

Nowak, M. and Sigmund, C. [1990] "The evolution of stochastic strategies in the prisoner's dilemma," *Acta Applicandae Mathematicae* **20**, 247−265.

Numaoka, C. [1998] "Personality development through interactions in virtual worlds," *Emotional and Intelligent: The Tangled Knot of Cognition, Papers from the AAAI Fall Symposium, Technical Report FS-98-03* (AAAI Press, Menlo Park, California), pp. 135−140.

Pessoa, L. [2008] "On the relationship between emotion and cognition," *Nature Reviews Neuroscience* **9**(2), 148−158.

Pfeifer, R. [1994] "The fungus eater approach to emotion: A view from artificial intelligence," *Cognitive Studies* **1**, 42−57.

Picard, R. W. [1995] Affective computing, Technical Report 321, M.I.T. Media Laboratory; Perceptual Computing Section.

Picard, R. W. and Scheirer, J. [2001] "The galvactivator: A glove that senses and communicates skin conductivity," *Proceedings from the 9th International Conference on Human-Computer Interaction*, New Orleans, pp. 1538−1542.

Picard, R. W., Vyzas, E. and Healey, J. [2001] "Towards machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(10), 1175−1191.

Poundstone, W. [1993] *Prisoner's Dilemma* (Doubleday, New York).

Reilly, W. S. [1996] *Believable Social and Emotional Agents*, PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Technical Report CMU-CS-96-138.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S. and Kilts, C. D. [2002] "A neural basis for social cooperation," *Neuron* **35**(2), 395−405.

Rilling, J. K., King-Casas, B. and Sanfey, A. G. [2008] "The neurobiology of social decision-making," *Current Opinion in Neurobiology* **18**, 159−165.

Scheutz, M. [2001] "The evolution of simple affective states in multi-agent environments," *Emotional and Intelligent II: The Tangled Knot of Social Cognition, Papers from the AAAI Fall Symposium, Technical Report FS-01-02* (AAAI Press, Menlo Park, California), pp. 123−128.

Simon, H. A. [1967] "Motivational and emotional controls of cognition," *Psychological Review* **74**(1), 29−39.

Sloman, A. [1998] "Damásio, Descartes, alarms and meta-management," *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2652−2657.

Sloman, A. [1999] "Review of affective computing," *AI Magazine* **20**(1), 127−133.

Staller, A. and Petta, P. [1998] "Towards a tractable appraisal-based architecture," *Workshop: Grounding Emotions in Adaptive Systems* (SAB'98: From Animals to Animats), pp. 56−61.

Toda, M. [1982] "Emotional Fungus-eaters," in *Man, Robot and Society* (The Hague: Nijhoff), pp. 130−153.

Tomlinson, B., Downie, M., Berlin, M., Gray, J., Lyons, D., Cochran, J. and Blumberg, B. [2002] "Leashing the AlphaWolves: Mixing user direction with autonomous emotion in a pack of semi-autonomous virtual characters," *Symposium on Computer Animation, Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (ACM Press, San Antonio, Texas), pp. 7−14.

Trivers, R. L. [1971] "The evolution of reciprocal altruism," *The Quarterly Review of Biology* **46**(1), 35−57.

Velásquez, J. D. [1998] "Modeling emotion-based decision-making," *Emotional and Intelligent: The Tangled Knot of Cognition, Papers from the AAAI Fall Symposium, Technical Report FS-98-03* (AAAI Press, Menlo Park, California), pp. 164−169.

Ventura, R. [2008] *Emotion-Based Mechanisms for Decision Making in Autonomous Agents*, Ph.D. thesis, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal.

Ventura, R., Custódio, L. and Pinto-Ferreira, C. [1998] "Emotions — the missing link?" *Emotional and Intelligent: The Tangled Knot of Cognition, Papers from the 1998 AAAI Fall Symposium, Technical Report FS-98-03* (AAAI Press, Menlo Park, California), pp. 170−175.

Ventura, R. and Pinto-Ferreira, C. [1998] "Emotion-based agents," *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)* (AAAI Press and MIT Press, Cambridge/Menlo Park), p. 1204.

Ventura, R. and Pinto-Ferreira, C. [2007] "Metric adaptation and representation upgrade in an emotion-based agent model," *Affective Computing and Intelligent Interaction: Second International Conference (ACII2007)*, LNCS 4738 (Springer), pp. 731−732.

Ventura, R. and Pinto-Ferreira, C. [2008] "Responding efficiently to relevant stimuli using an emotion-based agent architecture," *Neurocomputing* **72**(13−15), 2923−2930.

von Neumann, J. and Morgenstern, O. [1944] *Theory of Games and Economic Behavior* (Princeton University Press).

Vyzas, E. and Picard, R. W. [1998] "Affective pattern classification," *Emotional and Intelligent: The Tangled Knot of Cognition, Papers from the AAAI Fall Symposium, Technical Report FS-98-03* (AAAI Press, Menlo Park, California), pp. 176−182.

Wilson, I. [2000] "The artificial emotion engine, driving emotional behavior," *Artificial Intelligence and Interactive Entertainment, Papers from 2000 AAAI Spring Symposium, Technical Report SS-00-02* (AAAI Press, Menlo Park, California), pp. 76−80.

Wright, I. P. [1997] *Emotional Agents*, Ph.D. thesis, Faculty of Science of the University of Birmingham.